中 文 信 息 学 报
第 17 卷 第 3 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 17 No. 3

文章编号:1003 - 0077(2003)03 - 0053 - 06

汉语基调的调模与语音合成的质量提高*

吴禀雅[1],周昌乐[2],吴洁敏[3]

(1. 浙江大学,杭州 310028;2. 浙江大学人工智能研究所,杭州 310028;3. 浙江大学文学院,杭州 310028)

摘要:本文根据输入的汉语语篇中各个语词的感情色彩属性和语体色彩属性,通过一种语词属性文法及其合一运算,来得到整个语篇的调模。并通过调模得到相对应的音高和音长的基准值,来调整机器合成语音的语阶和语速,从而使机器合成的语音更加自然、流畅,丰富了机器合成语音的表现力,提高了语音合成的质量。

关键词:人工智能;机器翻译;调模;属性文法和合一运算;语速和语阶;质量提高

中图分类号:TP391.4 **文献标识码**:A

The Tonal Template of Mandarin Basic Tune and Quality Improvement of the Machine Speech Synthesis

WU Bing-ya[1], ZHOU Chang-le[2], WU Jie-min[3]

(1. Zhejiang University 310028; 2. Zhejiang University 310028; 3. Zhejiang University 310028, China)

Abstract: This paper achieves the tonal template of Mandarin discourse, by judging from the emotional color and colloquial style of every word in the input discourse and using a kind of phraseological attribute grammar and the relevant combinational arithmetic. The speed and scale of the machine synthesized speech are adjusted by the basic value of the pitch and duration of syllables corresponding to the tonal template, therefore the naturalness and fluency of the machine synthesized speech are enhanced, i.e. the quality improvement is realized.

Key words: artificial intelligence; machine translation; tonal template; attribute grammar and combinational arithmetic; speed and scale of speech; quality improvement

一、引言

语音合成,也即文语转换系统 TTS(Text To Speech),是将输入计算机的文本转换成语音形式输出的系统。近几十年来,该领域的研究取得了较大的进步。二十世纪七八十年代,主要采用的是参数合成的方法,比如 Holmes 的并联共振峰合成器和 Klatt 的串/并联共振峰合成器,通过选择合适的参数,来合成自然的语音。到二十世纪九十年代,较常用的是利用时域波形拼接方法来合成语音,它避免了参数合成法中提取参数的困难,合成语音的自然度较高。近期,还有一些新的合成技术陆续出现。如基于数据库的合成方法,语音单元从一个语音数据库中选择并拼接而成合成语句,清晰度和自然度都较高。

一直以来,语音合成的质量问题都是机器语音处理领域的重要研究内容。近年来许多优

* 收稿日期:2003 - 02 - 17
基金项目:浙江省教育厅科研资助项目(20010151)
作者简介:吴禀雅(1973 —),女,浙江金融职业学院讲师,主要研究方向为人工智能、计算语言学.

秀的 TTS 软件,如中科院声学所的 KX-PSOLA、清华大学的 TH-SPEECH、联想佳音、中国科技大学的 KDTALK,还有微软的 Microsoft Speech SDK 等,都较好地实现了输入语篇后能即时转换出语音的任务。而且这些系统合成的汉语语音的可懂度、清晰度都达到了较高的水平。

但这些技术和方法虽然在一定程度上提高了合成语音的自然度和清晰度,可是由于都没有将讲话的风格和感情色彩与具体的声学参数联系起来,所以合成的语音仍然缺乏情感表现力和自然度。而为了让机器合成语音能基本达到日常语言交流的要求,情感的表现力是非常重要的。为了表达说话者的感情,合成语音的语气、语速等都应该表现出一定的特征。而相对来说,利用波形拼接的方法来加强合成语音的表现力比较困难,而利用参数合成法,选择合适的参数来控制语气和语速,可行性更高。但具体参数的获取是主要问题所在[1]。

我们发现,长期以来,在汉语机器理解和语音合成研究方面,一直缺乏对汉语节律方面的研究利用。这当然不是因为汉语节律在汉语理解和语音合成中不重要,而是因为汉语节律的研究和利用,是一个更为深入的课题。需要在大量有关汉语韵律、节奏和语调等研究成果的基础上,才能够加以利用,并且通过形式化方法,来进行汉语节律的机器自动分析研究。

而吴洁敏、朱宏达教授发表的《汉语节律学》一书,恰好为我们提供了非常完整的汉语节律学的研究成果。在这个基础上,我们采用了意群动力学与属性文法合一运算相结合的方法,来进行汉语节律的机器自动分析研究,并尝试找到了用具体的声学参数来表现语篇的感情色彩和语体色彩的方法,进而提高语音机器合成的质量,使得机器合成的语音更富于感情。

二、汉语节律学

节律又叫声律、音律、韵律、上加成素或超音质成分、超音段音位、超音段特征、非线性特征等等。研究汉语的节律特征及其组合规律的科学叫汉语节律学,也可叫做节律语音学或语音修辞学。汉语节律学要研究的就是从音节组成音步、气群、句子、段落,乃至语篇的各个层级中的节律特征及其组合规律[2]。

汉语节律具有表意、表情、表态的功能。附着在音节和音节组合中的声音的高低、轻重、长短、快慢、停延和音质变化造成了语言的节律;同样内容的词句,采用不同的语气、间歇、长短、快慢,效果可能完全不同。而计算机的语音合成,一般都是一字一音,不但读不出作品的思想感情,甚至会由于停延位置不当、声调无正确处理等方面的原因,表达不出语篇的正确意思。

要寻求作品的神韵、读出作品的感情,必须根据作品内容找到语篇基调,然后根据基调来确定调模,再依此读出停延、节奏、重音和句调,就自然产生"以声传情"的作用。电脑在语音合成时,如果没有掌握好这些规律,就会产生怪腔怪调。

三、汉语语篇基调和调模

语篇基调属语篇层节律特征,也就是语篇的基本腔调,包括调型和调模两个方面。

其中的调模包括语阶和语速两个方面,也就是指语篇的音高、音长变化模式。调模具有表意功能。

调模中的语阶部分是由句外的音高变化形成的。篇章基调的语阶,从最低到最高,分为低语阶(A)、中语阶(B)、高语阶(C)三个等级。若同一句话的调型相同,而语阶不同,则表示的内容也许就不一样。语速由音步时值的变化形成。篇章基调的语速,从最快到最慢,分为快语速(a)、中语速(b)、慢语速(c)三个等级。语速不同,表示的意思不同。

构成语篇基调的主要成素有音高、音长和音强,但音强已经失去其区别词义的作用,可以

暂时不考虑音强特征,只求其音高和音长的二维参数。这里可根据吴洁敏教授提出的"汉语基调的九宫调模矩阵",把汉语基调的音高和音长作为两个坐标,音长作竖标,从低到高分为低语阶(A)、中语阶(B)、高语阶(C)三档;把音长作横标,从快到慢分为快语速(a)、中语速(b)、慢语速(c)三档。用虚线把三档语阶和语速连起来,就有了基调调域图上的低快调模 Aa、低中调模 Ab、低慢调模 Ac,和中快调模 Ba、中中调模 Bb、中慢调模 Bc,以及高快调模 Ca、高中调模 Cb、高慢调模 Cc,共九种基调模式。

图 1　汉语基调的九宫调模矩阵

高快 Ca	高中 Cb	高慢 Cc
中快 Ba	中中 Bb	中慢 Bc
低快 Aa	低中 Ab	低慢 Ac

汉语九宫调模矩阵可以涵盖所有语篇基调。如,高语阶、快语速的极端调模 Ca,可以表示激动或着急;高语阶、中语速的二维调模 Cb,可用于慷慨陈词的演讲中;高语阶、慢语速的极端调模 Cc 表示特殊感情,较少用;中语阶在日常生活中用得最多,中语阶、快语速 Ba 调模也较常见;中语阶、中语速的 Bb 调模平时用得最多,通常用于报告、讲课以及工作和生活中的正式谈话体;中语阶、慢语速的 Bc 调模大多用在文艺体中;低语阶、快语速的二维调模 Aa 在基调调模系统中,是较少见的;低语阶、中语速调模 Ab 在近距离的口语体中较常用;低语阶、慢语速的二维调模 Ac 在文艺体中较常见。

四、调模的获得

调模包括语阶和语速两个方面,是语篇的音高和音长变化模式。根据九宫调模矩阵,汉语基调共有九种模式,采用哪一种就要由我们的语篇内容决定。

我们知道,汉语中的词除了概念义以外,还有附属义。而词的附属义中,又包括词的感情色彩和语体色彩。

根据符淮青《现代汉语词汇》一书中所述,词的感情色彩一般有褒义、中性和贬义。词意为肯定的词语,如温顺、淳朴、壮实、漂亮、雄伟、贡献、英雄、珍品等,表示说话人对有关事物的赞许、表扬,感情色彩为褒,称为"褒义词";词意为否定的词语,如凶残、蛮横、丑陋、卖命、勾结、败类等,表示说话人对事物的厌恶、批评,感情色彩为贬,称为"贬义词";还有许多词语不带固定的感情色彩,称为"中性词"[3]。

词的语体色彩指不同的词适用于社会交际的不同范围,适用于不同文体的情况。词语根据其语体色彩可以分为两大类:书面语和口头语。书面语词用于书面写作,口头语词用于日常谈话。如吓唬、小气、[illegible]达、睡觉为口头语词,而与之意思相同的恐吓、吝啬、散步、睡眠则为书面语词。还有一些词语在口语和书面语中都能通用、没有明显区分的,我们称其为中性词。

根据我们平时收听各电视、电台播音员、主持人等的播音,及我们自己平时交谈的经验,我们不难分辨出:一般褒义词我们会说得响亮、高兴;而贬义词我们会说得低沉一些。在朗读书面语时,我们一般会语速较慢,咬文嚼字;但在口头说话时,我们一般都说得比较快。这样,我们刚好也将词的两种附属义作为两个坐标,词的感情色彩作竖标,从低到高分为贬义词、中性词和褒义词三档;把词的语体色彩作为横标,从左到右分为口头语、中性语和书面语三档。用虚线把三档词的感情色彩和语体色彩连起来,就有了和九宫调模矩阵相对应的矩阵,并且处于各个位置的词就采用九宫调模矩阵中对应位置的语阶和语速。

也就是说,当我们知道了语词的感情色彩属性和语体色彩属性后,我们就可以对照图 2,得到它的调模。

图 2　词的感情、语体色彩与调模对应图

词的感情色彩	口头语	中性语	书面语
褒义词	高快 Ca	高中 Cb	高慢 Cc
中性词	中快 Ba	中中 Bb	中慢 Bc
贬义词	低快 Aa	低中 Ab	低慢 Ac

词的语体色彩

但是我们知道,一篇语篇中有多个词语,而且各个词语的感情色彩属性和语体色彩属性都不一定全部相同。那么,如何根据语篇中的各个词语的感情色彩属性和语体色彩属性,来得到整个语篇的感情色彩属性和语体色彩属性,并从而最终决定整个语篇的调模呢?这里我们可以通过一种简单的语词属性文法及其合一运算来实现。

五、算法实现

输入汉语语篇后,首先我们要进行语词切分[4]。然后分别标注出每个词语的感情色彩属性值 emotion(i)和语体色彩属性值 style(i),其中 i 表示语篇中的词语序号。emotion(i)的值为 1～3,emotion(i) = 1 表示该词为贬义词,emotion(i) = 2 表示该词为中性词,emotion(i) = 3 表示该词为褒义词。style(i)的值也为 1～3,style(i) = 1 表示该词为口头语,style(i) = 2 表示该词为中性语,style(i) = 3 表示该词为书面语。

知道语篇中每个词语 i 的感情色彩属性值 emotion(i)和语体色彩属性值 style(i)后,整个语篇的感情色彩属性和语体色彩属性标注就可以分别通过以下运算公式来进行:

$$P_j = \frac{1}{n}\sum_{i=1}^{n} P_j(i)$$

其中,$j = 1, 2$。P_1 为语篇的情感色彩属性,P_2 为语篇的语体色彩属性。n 为语篇中的词语数,$P_j(i)$ 为第 i 个词语的属性,当该词语为"褒义",也即 emotion(i) = 3,或为"口头语",也即 style(i) = 1 时,$P_j(i)$ 的值取为 + 1;当该词语为"中性词"时,也即 emotion(i) = 2 或 style(i) = 2 时,$P_j(i)$ 的值取为 0;当该词语为"贬义",也即 emotion(i) = 1,或为"书面语",也即 style(i) = 3 时,$P_j(i)$ 的值取为 - 1。

通过该公式的层层合一,就可以给出整个语篇的感情色彩属性和语体色彩属性的具体值。

我们可以定义:当 $P_1 > 0$ 时,该语篇的情感色彩属性为褒义;当 $P_1 = 0$ 时,该语篇的情感色彩属性为中性;当 $P_1 < 0$ 时,该语篇的情感色彩属性为贬义。当 $P_2 > 0$ 时,该语篇的语体色彩属性为口头语;当 $P_2 = 0$ 时,该语篇的语体色彩属性为中性;当 $P_2 < 0$ 时,该语篇的语体色彩属性为书面语。

这样,得到了整个语篇的情感色彩属性和语体色彩属性后,再对照图 2,我们也就最终可以得到语篇的调模。

另外,根据吴洁敏教授的语音实验的结果,汉语基调的三级语阶的参数如表 1 所示。而汉语基调的三级语速的参数如表 2 所示。

表 1　汉语基调的三级语阶参数

语阶的音高级	基调上限 —下限参数
高语阶(C)	450～270(Hz)
中语阶(B)	330～150(Hz)
低语阶(A)	250～110(Hz)

表 2　汉语基调的三级语速参数

语速级	音节平均时值参数
快速(a)	135～300ms/音节
中速(b)	250～450ms/音节
慢速(c)	400～600～1000ms/音节

这样,根据刚才通过合一算法所得到的语篇的调模,再参照表 1 和表 2,我们就可以得到整个语篇的音高和音长的基准值了。

六、实验与结果分析

例句:“打从今天起,大伙要珍惜这宝贵的机会,多为班级建设出力!”

我们利用 Microsoft Speech SDK 5.0 的 TTSApp 软件,试着朗读出了该例句,并得到了该例句的输出波形:

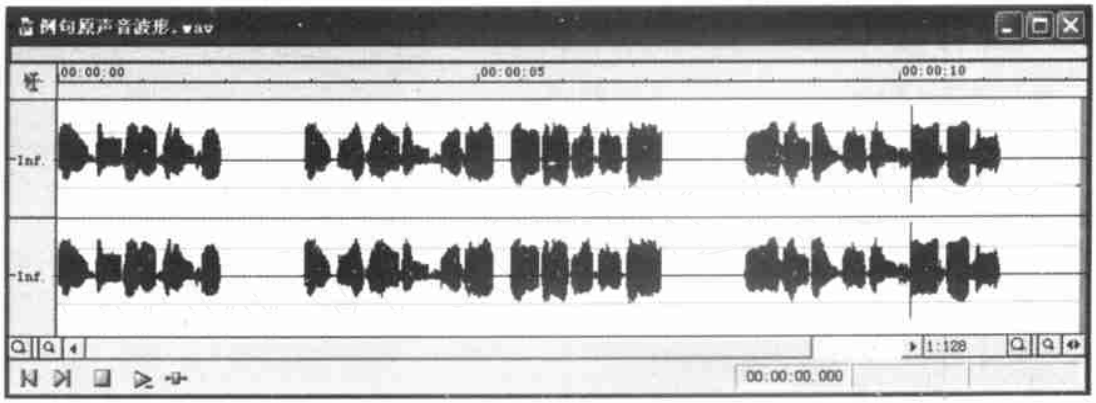

图 3 例句原声音波形图

接着我们利用北京大学计算语言学研究所的“汉语文本切分与词性标注(Chinese Text Segmentation and POS Tagging)”系统,来进行语词的切分和词性的标注,结果为:

打/v 从/p 今天/t 起/v ,/w 大伙/r 要/v 珍惜/v 这/r 宝贵/a 的/u 机会/n ,/w 多/a 为/u 班级/n 建设/v 出力/v !/w

其中的词类标记,请参看北大计算语言学研究所的汉语文本词性标注标记集。(http://icl.pku.edu.cn)

我们查到各个词语的感情色彩属性为:

打/中 从/中 今天/中 起/中 ,大伙/中 要/中 珍惜/中 这/中 宝贵/褒 的/中 机会/中 ,多/中 为/中 班级/中 建设/中 出力/中 !

其中:/中指中性词,/褒指褒义词,/贬指贬义词。

各个词语的语体色彩属性为:

打/口 从/中 今天/中 起/中 ,大伙/口 要/中 珍惜/中 这/中 宝贵/中 的/中 机会/中 ,多/中 为/中 班级/中 建设/中 出力/中 !

其中:/中指中性词,/口指口头语,/书指书面语[5]。

通过语词属性标注和公式: $P_j = \frac{1}{n}\sum_{i=1}^{n} P_j(i)$

我们计算出语篇的情感色彩属性 P_1 为 $\frac{1}{16}$,语体色彩属性 P_2 为 $\frac{2}{16}$。从而得到语篇的调模为 Ca。再通过查阅表 1 和表 2,我们可以得到语篇基调的上限—下限参数为 450Hz～270Hz,音节平均时值参数为 135～300ms/音节。

根据得到的语篇基调的音高和音长值,我们利用 Sonic Foundry, Inc. 的 Sonic Foundry Sound Forge 6.0 软件,对原声音波形进行了处理,使其符合我们的要求:其中的每个音节的时长从原来的大约 389ms,变为修改后的大约平均每音节时长 233ms;而频率则调整为 450Hz 左右。

处理后的声音波形如下:

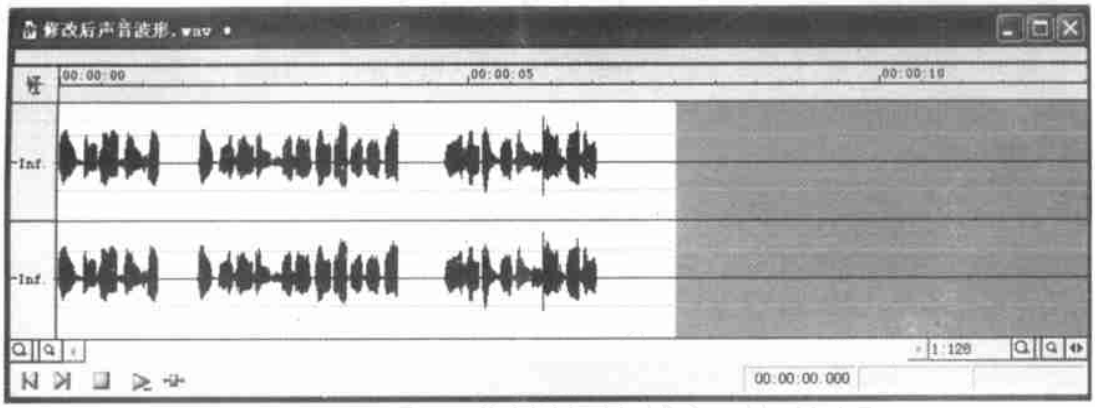

图 4　修改后例句声音波形图

我们尝试将原始的机器合成语音和修改后的语音播放给 100 位大学生听,并请他们比较两者的效果,结果认为后者表达出了积极、激动的情感的占 62 %,有 56 %的大学生觉得听起来后者是在口头说话;认为后者比前者自然的占 90 %,觉得后者的语音质量比前者要高的占 84 %。

这说明根据语篇调模来对机器合成语音进行调整后,丰富了语音的情感表现力,语音的自然度得到了改善,语音合成的质量是有明显提高的。

七、结论和不足之处

本文在《汉语节律学》的研究成果的基础上,先对汉语语篇进行了语词切分,再根据语篇中各个语词的感情色彩属性和语体色彩属性,通过一种语词属性文法及其合一运算,得到了整个语篇的情感色彩属性和语体色彩属性。以此对照汉语基调的九宫调模矩阵,得到了整个语篇的调模。通过与该语篇调模相对应的音高和音长的基准值,来调整机器合成语音的语阶和语速,从而使机器合成的语音不再那么单调、平板,不再是不管什么内容都采用同样的语速和音高了,而是变得更加自然、流畅,更具情感的表现力。在没有大大增加语音合成系统的复杂度的前提下,提高了语音合成的质量。

但本文仅分析了汉语基调中的调模部分,对于汉语节律中其它的重要部分,如停延、重音、节奏和句调等都没有考虑,这制约了语音合成的自然度和质量提高的程度。在以后的研究中,如再加入这些因素,相信语音合成在质量上会有更大的提高,在准确度和自然度上会有更佳的表现。另外,本论文对于从调模得到语音频率和语速的范围后,到底采用哪个固定值未做深入研究,未能较好体现语篇感情和语体色彩强弱程度的差异,需要在以后进一步进行研究和探讨。

参　考　文　献:

[1]　初敏. 韵律研究与合成语音的自然度. 第五届现代语音学学术会议文集, 2001.
[2]　吴洁敏,朱宏达. 汉语节律学,北京:语文出版社,2001.
[3]　符淮青. 现代汉语词汇,北京:北京大学出版社,1985.
[4]　周强. 规则和统计相结合的汉语词类标注方法. 中文信息学报,1995(9):1 - 10
[5]　吕叔湘. 现代汉语八百词. 北京:商务印书馆,1981.